© THE QUEEN'S PRINTER FOR ONTARIO
1999

REPRODUCED WITH PERMISSION

L'IMPRIMEUR DE LA REINE POUR L'ONTARIO

REPRODUIT AVEC PERMISSION

micromedia
a division of IHS Canada

20 Victoria Street
Toronto, Ontario M5C 2N8
Tel: (416) 362-5211
Toll free: 1-800-387-2689
Fax: (416) 362-6161
Email: info@micromedia.on.ca

Education
Quality and
Accountability
Office

# EQAO Research Series

• NUMBER 3 • AUGUST 1999

100-04338/

## EQAO MISSION

EQAO will ensure greater accountability and contribute to the enhancement of the quality of education in Ontario. This will be done through assessments and reviews based on objective, reliable and relevant information and the timely release of that information along with recommendations for system improvement.

*This paper was written under a contract with the Education Quality and Accountability Office (EQAO). The opinions expressed are those of the authors and not necessarily those of EQAO.*

© Copyright 1999 Queen's Printer for Ontario

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, or otherwise, without the prior express permission of the Education Quality and Accountability Office's Manager of Communications and Marketing.

Education Quality and Accountability Office
2 Carlton Street, Suite 1200
Toronto, Ontario M5B 2M9

Telephone: 1-888-327-7377

Web site: www.eqao.com

ISBN 0-7778-9136-0

# Foreword

EQAO believes that large-scale assessment can contribute to positive educational change when it engages educators, parents and students in thought and discussion about what takes place in the classroom.

EQAO's mandate reflects two key commitments: accountability and improvement. We report to parents, educators and the public about student achievement and education quality in Ontario's publicly funded school system. We also work to ensure that this information serves as a catalyst for improvement in learning and teaching.

One of the ways in which we promote improvement is by supporting research into the factors that influence student achievement and education quality. The *EQAO Research Series* will provide information and analysis on a wide range of educational issues.

This series has two important purposes. The first purpose is to increase our knowledge about what works in classrooms, schools and school systems. The second purpose is to close the gap between what we know and what we do. Research can have a positive and lasting impact on student learning when it gives educators the knowledge, opportunities and motivation to adapt and refine their strategies and approaches based on accurate and reliable data. EQAO supports all forms of systematic inquiry that promote constructive change and improvement.

I am pleased to introduce this latest paper in the *EQAO Research Series.*

Joan M. Green
Chief Executive Officer

August 1999

## TABLE OF CONTENTS


Abstract ..... 1

Methodological Framework

- Generalizability Theory ..... 2
- Classification and Misclassification ..... 4

Implementation of Classification and Misclassification Analysis in EQAO Assessments ..... 6

- Data Collection and Summarization ..... 6
- The Statistical Model ..... 6
- Results ..... 9

Discussion and Conclusion ..... 11

References ..... 12

The Authors ..... 13

# Psychometric Perspectives for EQAO: Generalizability Theory and Applications

*Richard Wolfe*
*Ontario Institute for Studies in Education*
*University of Toronto*

*David Wiley*
*Northwestern University*

*Ross Traub*
*Ontario Institute for Studies in Education*
*University of Toronto*

*This paper examines the psychometric perspectives for defining and evaluating the generalizability of the measurements and reports produced by the Education Quality and Accountability Office (EQAO). Measurement accuracy is conceptualized in terms of the components of measurement variability: tasks, individuals, markers and related interactions. The analysis of the measurement generalizability of the EQAO assessments uses two psychometric perspectives. The first is generalizability analysis of the sources of variation in test scores (that is, components of true and error variance); the second involves models for true categorization of student performance and a probability framework for defining correct classification or misclassification. Data from the EQAO 1997–1998 French-language and English-language assessments in Grades 3 and 9 are examined by means of a latent class analysis methodology that estimates the rates of correct classification and misclassification for markers. The probability of correct classification for these assessments is estimated to be from 70% to 80%. This outcome is reasonably high and is comparable to what is observed in other large-scale performance assessments.*

# Methodological Framework

## Generalizability Theory

The most comprehensive approach taken by the educational measurement community in analyzing errors in student scores on performance exercises has been based on generalizability theory. This provides an expansion and extension of the methods used in traditional reliability theory in psychometrics, such as test-retest and inter-rater reliability coefficients. It builds heavily on the notion of the sampling of test tasks and measurement conditions from larger populations of such tasks and conditions.

Generalizability theory considers measurement reliability from the perspective of generalizing from such samples to the relevant populations. The populations are characterized by parameters representing variability in measured test performance. These parameters characterize two kinds of variability:

1. variability across individuals in performance characteristics that is commonly reflected across varying tasks and over varying conditions, that is, general variation in individual performance that transcends particular circumstances of measurement, and
2. specific variation in individual performance that reflects only particular tasks and conditions.

The former are desired measurement stabilities, whereas the latter are viewed as measurement errors. Generalizability studies and analyses assess the magnitude of these aspects of performance variation by conducting analyses of variance. These analyses break down the observed variation in performance into variance components representing persons, tasks and other measurement conditions. The relative and absolute magnitudes of the different variance components can predict the accuracy that would be obtained with alternative measurement designs that use different numbers and conditions of the tasks being studied. Generalizability theory was first formulated by Cronbach, Gleser, Nanda and Rajaratnam (1972).

In performance assessments, measurement errors are most commonly associated with tasks and scorings. That is, they appear in variations in an individual's score on similar tasks and in different scorings. Specifically, these errors are assessed relative to score variation in task performance

- across different performance exercises regarded as interchangeable for purposes of the assessment, and
- across different scorings of a single task also regarded as interchangeable for purposes of assessment.

The greater these variations, the greater the measurement error and the lower the accuracy. In most performance assessments, exercises scored in a particular scoring context within a particular subject and grade are usually regarded as interchangeable, particularly when the scorers are trained in the same way to the same level of proficiency.

Data for a generalizability analysis of these kinds of errors are often obtained from a situation in which each of a group of test takers performs every task in a constant set and each response is scored by the same set of scorers. An analysis of the data with this structure, when there is more than one task in the set, yields the estimates of components of variability given in Table 1.

All the types of error listed in Table 1 contribute to error. This is true regardless of the number of tasks completed by a respondent in a set of interchangeable tasks or the number of interchangeable scorings given to a completed performance. However, depending on the assessment design, e.g., the actual number of tasks completed by a respondent or the number of markers, the effects of the different components will vary. If only one task of a given type is completed by respondents or if only one scoring is obtained for a completed task, some of these components will not be estimable.

## Methodological Framework

### Generalizability Theory

Most types of error can be traced to two general sources: differences in the measuring properties of tasks and differences in the scorers' assessments of task performances. Since tasks are selected, and hence fixed, before scoring takes place, task differences can be conceived as prior to and independent of scoring effects. On the other hand, scoring effects can take place only in the context of the tasks given. From this perspective, it makes sense to define error components coming about from task differences (task effects) as those involving task but not involving scoring, and to define scoring effects as those involving scoring, whether or not they involve task. That is, the results for a particular task will usually differ randomly from those that would be found after an assessment of performance on a large collection of interchangeable tasks. Scoring, on the other hand, is task-specific: scorers score only specific tasks. From this viewpoint, the error variance components in Table 1 would be categorized as shown in Table 2.

Error due to task differences in performance, exclusive of scoring, is reflected in reliability as

$r_t = \sigma^2_p / ( \sigma^2_p + \sigma^2_t + \sigma^2_{pt})$.

Error due to scoring, once the tasks are fixed (so that the task-specific true score is the observed score after error is added), is reflected in reliability as

$r_s = (\sigma^2_p + \sigma^2_t + \sigma^2_{pt}) / (\sigma^2_p + \sigma^2_t + \sigma^2_{pt} + \sigma^2_s + \sigma^2_{ps} + \sigma^2_{st} + \sigma^2_{pst})$.

Since the total reliability, reflecting both sources of error, is

$r_{tot} = \sigma_p^2 / (\sigma_p^2 + \sigma_t^2 + \sigma_{pt}^2 + \sigma_s^2 + \sigma_{ps}^2 + \sigma_{st}^2 + \sigma_{pst}^2)$,

then

$r_{tot} = r_t r_s$.

That is, total reliability is the product of task error reliability and scoring error reliability.

Although all the above components are present in the EQAO assessment process, the current design of the assessments limits which of them can be evaluated. Future research studies can be designed to identify components that could feasibly be incorporated into the main data collection stage and included in subsequent analysis.

**TABLE 1 Types of Variation**

| Variance of Component | Type of Variation |
|---|---|
| True Variation | |
| $\sigma^2_p$ | General "true" variation across persons in performance |
| Error Variation | |
| $\sigma^2_s$ | Variation across scorers in average leniency |
| $\sigma^2_t$ | Variation across tasks in average difficulty |
| $\sigma^2_{ps}$ | Variation across scorers and persons in leniency toward individuals |
| $\sigma^2_{pt}$ | Variation across tasks and persons in task difficulty for individuals |
| $\sigma^2_{st}$ | Variation across scorers and tasks in scored task difficulty |
| $\sigma^2_{pst}$ | Variation across scorers, tasks and persons in task difficulty and leniency toward individuals |

**TABLE 2 Error Components for Task and Scoring**

| Source | Error Components |
|---|---|
| Task | $\sigma^2_t$, $\sigma^2_{pt}$ |
| Scoring | $\sigma^2_s$ , $\sigma^2_{ps}$ , $\sigma^2_{st}$ , $\sigma^2_{pst}$ |

## Methodological Framework

### Classification and Misclassification

The generalizability approach to the analysis of measurement error treats the responses analyzed as if they could be assigned values on a continuous scale. An alternative approach assigns the responses to one of several discrete, ordered categories. From this perspective, measurement error consists of the assignment of scores to incorrect categories. The natural way of quantifying these misclassifications is in terms of probability of correct classification or misclassification, that is, probabilities such as P(ilj) = the probability of a response being assigned to category i when j is actually the correct category.

Misclassification can occur because of any of the types of error listed in Tables 1 and 2. In particular, if an individual were truly in category j, a task t might have probability P(ilj;t) of categorizing that person in category i. This task-specific misclassification probability would arise from the same sources summarized by $\sigma^2_t$ and $\sigma^2_{pt}$. Also, an average of these misclassification probabilities over tasks, that is, $\text{mean}_t\{P(ilj;t)\}$, summarizes the general relationship across tasks between true and observed categories.

Misclassification relations can be arrayed as a misclassification matrix, as illustrated in Table 3.

These matrices convert true into observed values. That is, if $y_p$ is a vector representing a categorization of individual p into an observed category (that is, the vector element = 1 for the observed category, and 0 for each other category), and if $x_p$ is the corresponding true vector, then $E_m(y_p) = Mx_p$, where $E_m$ is the expectation over the misclassification process (task performance or scoring). Similarly, $E_p[E_m(y_p)] = E_p[Mx_p] = ME_p[x_p]$, where $E_p$ is the expectation over the population of individuals. Note that $E_p[x_p] = \pi$, the true proportions of the categories in the population, and $E_p[E_m(y_p)] = p$, the observed proportions of the categories in the population. Thus, $p = M\pi$. That is, misclassification matrices transform true distributions over categories for groups of respondents into observed distributions over categories across tasks or scorings. If we index these values for a specific task, say t, then $p_t = M_t\pi$.

If the misclassification due to scoring task t is $M_{ts}$, then $p_{ts} = M_{ts}p_t = M_{ts}M_t\pi$, so that the misclassification matrix of true performance after both task selection and scoring is the matrix product of the task selection misclassification and the scoring misclassification matrices. Note also that $\Sigma_t(p_{ts}) = \Sigma_t(M_{ts}M_t\pi) = [\Sigma_t(M_{ts}M_t)]\pi$, so that the sum and mean of the misclassifications over tasks are the misclassifications for the sum and mean of the observed proportions.

Each diagonal entry in the misclassification matrix represents the probability of correct classification, that is, that an individual's task in a particular true category will be scored in that category. These diagonal entries represent measurement accuracy or reliability. (Note that classical reliability indexes, as described above for generalizability theory, are correlation ratios, not probabilities. We find the probabilities of the latent class analysis more interpretable, especially where categorical scales are used as in the EQAO assessments. While the range of probabilities and correlation ratios is the same, zero to one, the scale is not, so the two are not directly comparable. We present a way of transforming classical reliabilities to probabilities later in this paper.)

## Methodological Framework

### Classification and Misclassification

**TABLE 3 Formulation of the Misclassification Matrix, M**

| | True Category | | | |
|---|---|---|---|---|
| **Observed Category** | 1 | 2 | 3 | 4 |
| 1 | P(1\|1) | P(1\|2) | P(1\|3) | P(1\|4) |
| 2 | P(2\|1) | P(2\|2) | P(2\|3) | P(2\|4) |
| 3 | P(3\|1) | P(3\|2) | P(3\|3) | P(3\|4) |
| 4 | P(4\|1) | P(4\|2) | P(4\|3) | P(4\|4) |
| **Total** | 1.00 | 1.00 | 1.00 | 1.00 |

# Implementation of Classification and Misclassification Analysis in EQAO Assessments

In the EQAO assessments in Grade 3 and Grade 9, we have to consider each student as having done one task, namely, the whole assessment procedure. Given EQAO's overall priorities and objectives, it was not feasible to have replicate tasks in any of the assessments, that is, separately scorable responses from the students that could be regarded as interchangeable. This means that the variance components that depend on task (task variance, task by student interaction, task by marker interaction and task by student by marker interaction) cannot be evaluated. So we must focus on the variances and reliabilities due to markers.[1]

## DATA COLLECTION AND SUMMARIZATION

The data for this analysis derive from a re-marking system that is used by EQAO. After about 500 to 1,000 papers have been marked, there is a quick analysis of the provisional mark distributions. A stratified sample of papers is selected with approximately uniform distribution over the scoring range. The overall or average mark is used in this selection. In the case of English papers for provincial assessments, in which there are more than 100,000 students, about 200 are chosen for calibration. For the French papers, and in the case of sample surveys (that is, the Grade 9 assessment), about 100 papers are chosen for calibration. Once selected, the papers are reinserted as often as possible into the marking flow, without visible signs that they are being repeated. Because there are so many papers and markers, it is very rare that a calibration paper is marked twice by any one marker. Depending on the speed of marking and the number of papers and markers, each calibration paper is re-marked from three to six times.

The re-marking data are summarized, separately for each assessment stream and for each mark that is to be analyzed, in a cross-tabulation. Examples of such cross-tabulations are given in Table 4. In Form A of the example, a randomly selected paper marked by two randomly selected markers has a 21.15% chance of being marked twice as level 2 and a 12.61% chance of being marked once at level 2 and once at level 3. The corresponding figures for Form B are 18.33% and 13.61%.

The overall distribution of marks can be determined from the cross-tabulation by adding the diagonals to half the sum of the off-diagonals (since the cross-tabulation records pairs of markings). The distributions are 14.98%, 34.95%, 33.60% and 16.43% for levels 1, 2, 3 and 4 for Form A; and 19.45%, 34.58%, 32.08% and 13.90% for Form B. Although the same number of papers for reinsertion was chosen from each of the four levels after the first markings of papers, there is some regression to the middle in the averaged markings; that is, the percentages for levels 1 and 4 are diminished while those for levels 2 and 3 are increased.

## THE STATISTICAL MODEL

The data summaries depend on the conjunction of mark distribution statistics and marker agreement rates. To disaggregate these effects, we apply a latent class model as illustrated in Figure 1. (For the original explanation of latent class analysis, see Lazarsfeld & Henry, 1968.)

To understand the latent class analysis, we need to consider that in any real application

[1] In all the EQAO assessments, there are separate scores for major categories or domains, and topics or strands. These provide differentiated evaluations of the students, but they are not alternative or interchangeable for purposes of determining reliability. In several assessments and in field trials, there have been "analytic" scorings in which partial marks were determined for the various questions and segments of the student work. These correlate well with the holistic marks. The multiple-choice and short-answer tests given to samples of students also correlated well with the holistic marks. While this augurs well for the reliability of the holistic marks (the correlations could not exist without reliability of at least that level), it is not possible to quantify reliability through such connections because the methods for this type of analysis have not yet been developed.

we do not know the "true" rating for any paper. We might try to derive that from a consensus of experts, but even then we are unlikely to obtain 100% agreement. Without a "true" classification, we cannot directly calculate the proportion of correct and incorrect classification. On the other hand, the proportions of agreement, as indicated in Table 4, depend on both the distribution of hypothetical true ratings (the latent classes of the papers) and the marker classification probabilities (see Table 3).

The latent class analysis proposes a simple model for the rating process. First, the papers have a true, although unknown, latent (or true) classification, common to all markers. Second, when a marker gives a mark to a paper, this will be a correct classification with a certain probability. With the complementary probability, the marker marks the paper up or down one level. At level 1 (the lowest level), there is considered to be only correct or upward marking, with down merged with correct, and at level 4 (the highest level), there is only correct or downward marking, with up merged with correct. At the other levels there are constant but differentiated probabilities for up and down.

The proportions of papers in the true categories are parameters of the model to be estimated. The probabilities of marking correctly, of marking up and of marking down are three other parameters. This set of parameters is estimated from the data summarization by the process of maximum likelihood. This means that, by means of a numeric iteration, parameter values are found that make the observed summary proportions as likely as possible.

Returning to Figure 1, a diagram of the model is shown in the top panel, with the parameter values for latent classes and classification or misclassification as estimated from the data in the bottom panel. The middle panel shows the expected value of the data given the parameters, and we can see that it is a close approximation to the data.

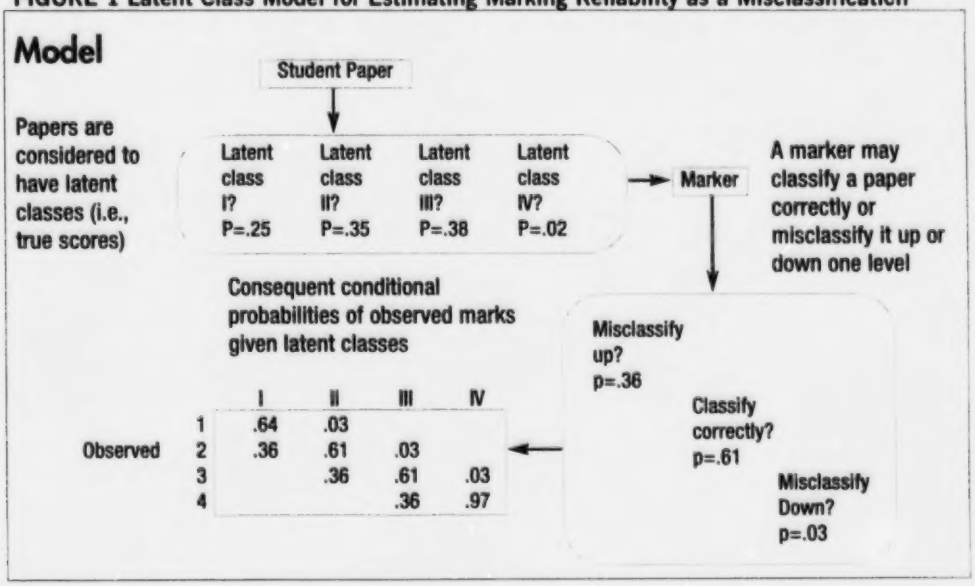
This particular latent class model is a special case of the model presented in the previous section. The simplification is necessary because of the relatively small sample sizes. The model seems to be adequate to fit the data, partly because there are very few cases where markers differ in their judgments by more than one level.

**TABLE 4 Example of Summaries of Classification and Misclassification Data**

| Form A Level | Level 1 | Level 2 | Level 3 | Level 4 |
|---|---|---|---|---|
| 1 | 7.70 | | | |
| 2 | 14.01 | 21.15 | | |
| 3 | 0.56 | 12.61 | 18.21 | |
| 4 | 0 | 0.98 | 17.60 | 7.14 |
| **Form B Level** | | | | |
| 1 | 10.28 | | | |
| 2 | 15.83 | 18.33 | | |
| 3 | 2.50 | 13.61 | 17.08 | |
| 4 | 0 | 3.06 | 13.89 | 5.42 |

*Note:* In each cross-tabulation the entries sum to 100% and estimate the distribution of marks as if a randomly selected paper is marked by two randomly selected markers.

**FIGURE 1 Latent Class Model for Estimating Marking Reliability as a Misclassification**


Model
Student Paper
Papers are considered to have latent classes (i.e., true scores)
Latent class I? P=.25
Latent class II? P=.35
Latent class III? P=.38
Latent class IV? P=.02
Marker
A marker may classify a paper correctly or misclassify it up or down one level
Consequent conditional probabilities of observed marks given latent classes
Observed
I II III IV
1 .64 .03
2 .36 .61 .03
3 .36 .61 .03
4 .36 .97
Misclassify up? p=.36
Classify correctly? p=.61
Misclassify Down? p=.03

**Expected Data**

| | I | II | III | IV |
|---|---|---|---|---|
| 1 | 7.7 | | | |
| 2 | 14.0 | 21.2 | | |
| 3 | .6 | 12.6 | 18.2 | |
| 4 | | 1.0 | 17.6 | 7.1 |

Theoretical average cross-tabulation, as a consequence of the parameters (sums to 100%)

**Observed Data**

| | I | II | III | IV |
|---|---|---|---|---|
| 1 | 10.1 | | | |
| 2 | 12.4 | 16.4 | | |
| 3 | .6 | 16.7 | 19.0 | |
| 4 | | .7 | 17.0 | 7.1 |

Observed average cross-tabulation, as calculated from the calibration data (sums to 100%)

*Note:* In the statistical analysis, we try to find model parameters for which the observed data are most likely, that is, closest to the expected data, given the parameters.

## RESULTS

The results from applying the latent class model to the 1997–1998 assessments are given in Tables 5 and 6. As pointed out earlier, we can regard the true classification probabilities as a good summary measure of reliability, that is, the probability that a student will be marked "correctly" by a marker. These probabilities are in the range 70% to 80%. That is reasonably high and is comparable to levels that are usually found in other large-scale performance assessments (see, for example, Dunbar, Koretz & Hoover, 1991). Furthermore, these classification rates are comparable to what would be obtained in traditional multiple-choice tests (see following section).

**TABLE 5 Latent Class Analysis for 1997–1998 Overall Marks: English-Language**

| Score | LC0 | LC1 | LC2 | LC3 | LC4 | Down | Up | True |
|---|---|---|---|---|---|---|---|---|
| English G3 Reading | .10 | .08 | .35 | .42 | .05 | .11 | .07 | .81 |
| English G3 Writing | .09 | .09 | .46 | .36 | .00 | .06 | .20 | .74 |
| English G3 Mathematics | .09 | .11 | .41 | .38 | .00 | .08 | .18 | .73 |
| English G9 Mathematics | .12 | .15 | .44 | .28 | .01 | .02 | .26 | .72 |

**TABLE 6 Latent Class Analysis for 1997–1998 Overall Marks: French-Language**

| Score | LC0 | LC1 | LC2 | LC3 | LC4 | Down | Up | True |
|---|---|---|---|---|---|---|---|---|
| French G3 Reading | .10 | .05 | .30 | .55 | .00 | .13 | .12 | .74 |
| French G3 Writing | .10 | .06 | .33 | .53 | .00 | .09 | .16 | .74 |
| French G3 Mathematics | .12 | .05 | .29 | .46 | .08 | .12 | .08 | .79 |
| French G9 Mathematics | .07 | .13 | .40 | .38 | .02 | .25 | .05 | .69 |

### Classification Rates in Conventional, Multiple-Choice Tests

For reasons of content and consequential validity, EQAO assessments use primarily performance tasks measured holistically. It is interesting to compare the reliabilities that might be obtained if traditional testing were used instead. Students could be categorized into, say, four levels (to parallel the EQAO performance markings) by dividing the score distribution on a conventionally scaled test, such as a multiple-choice test, at appropriate score thresholds. Using classical test reliability theory, we can estimate the probabilities of correct classification that would be obtained in such a procedure.

As an example, suppose that a scaled test resulted in the distribution of final levels similar to those in the EQAO 1997–1998 Grade 3 Assessment, with 10% in level 1, 40% in level 2, 45% in level 3 and 5% in level 4. (These represent an approximate average distribution over EQAO's subjects and languages.)

We will consider scaled tests with classical reliability coefficients (ρ) of .70, .80 and .90, which correspond to medium, high and very high accuracy for commercial educational tests in mathematics and language. For each test, we could calculate the probabilities of correct-level assignments for students who are at the centres of levels 1 to 4, as shown in Table 7.

For example, in Table 7, the figure 73 for level 2 with test reliability .70 says that the chances are 73% that a student at the true centre of level 2 on this test (which means in this case at the 30th percentile of the true score distribution) will be classified in level 2. (A more detailed analysis shows that the student has a 6% chance of being classified in level 1, a 21% chance of being classified in level 3, and practically zero chance of being classified in level 4.)

We see that the probabilities of correct classification are higher for levels 2 and 3, because these are broad categories and we are considering only students in their centres. For students closer to the score thresholds of the levels, the probabilities of correct classification would be lower. The chances of correct classification are also substantially lower in the outer levels, 1 and 4, because these levels are narrower and because there is general regression toward the centre.

The lesson to be learned is that conventional multiple-choice tests, even of very high reliability, will produce substantial classification errors when used to assign students to levels. In addition, the multiple-choice tests are considered to have significant defects in validity and bias. On the other hand, the rates of correct classification from the single marker as produced in the EQAO performance assessments are of comparable magnitude.

**TABLE 7 Probability of Correct Classification of Students at the Level Centre for Tests with Different Reliability Coefficients**

| Reliability | Level 1 | Level 2 | Level 3 | Level 4 |
|---|---|---|---|---|
| .70 | 57 | 73 | 80 | 50 |
| .80 | 66 | 82 | 88 | 60 |
| .90 | 81 | 94 | 96 | 75 |

# Discussion and Conclusion

Inter-task and inter-rater reliabilities are sometimes defined for this kind of testing and marking work. However, analysis of inter-task reliability is not feasible here, because there is just one task per student. For purposes of the present analysis, our focus therefore has been on inter-rater reliability. A sample of papers with uniform distribution over the scoring range was reinserted between three and six times into the marking flow. A paper was virtually never marked twice by the same marker. What was computed under the model presented here is the accuracy of the mark that was assigned to each paper as given by the probability of correct classification. (This calculation is clearly different from the probability that two markers coincide in their assignments. Though we could calculate the probability that two markers would coincide, under the model this probability would include cases where markers coincide in making errors.)

It is important to emphasize that the latent class distributions are not the same as the original distribution of marks. They give a presumably better picture of the distribution of the "true" classifications of these students, because the observed distributions are distorted by the misclassifications. Note that the sample of papers used in the calibration exercise was chosen to have a more or less uniform distribution over the range of marks, so the latent class proportions may not be reflective of the total population. In any case, one can see that the number of cases in latent class 4 is very small.

From these analyses, we can begin to form an evaluation of the sufficiency of reliability of the EQAO measurements and reports. Generally speaking, the probabilities of correct classification are reasonably high. These rates are comparable to what is found in other large-scale performance assessments, and they are equivalent to the classification rates that would be obtained in more traditional multiple-choice tests, which are usually considered to have very high reliability.

# References


Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability of scores and profiles.* New York: Wiley.

Dunbar, S.B., Koretz, D.M., & Hoover, H.D. (1991). Quality control in the development and use of performance assessments. *Applied Measurement in Education,* 4(4), 289-303.

Lazarsfeld, P. F., & Henry, N. W. (1968). *Latent structure analysis.* Boston: Houghton Mifflin.

# The Authors

Richard Wolfe is Associate Professor in the Measurement and Evaluation program of the Department of Curriculum, Teaching and Learning of the Ontario Institute for Studies in Education of the University of Toronto. His research areas include design and analysis of large-scale assessments.

David Wiley is Professor, on leave, in the Program on Learning Sciences and former Dean of the School of Education and Social Policy, Northwestern University. His research work has been in statistics and in curriculum, learning and testing theory.

Ross Traub is Professor in the Measurement and Evaluation program of the Department of Curriculum, Teaching and Learning of the Ontario Institute for Studies in Education of the University of Toronto. He has research interests in theoretical and applied psychometrics.

OQRE

Office de
la qualité et
de la responsabilité
en éducation

# OQRE Série d'études de recherche

• NUMÉRO 3 • AOÛT 1999

## MISSION DE L'OQRE

L'OQRE assurera une plus grande responsabilité et contribuera à améliorer la qualité de l'éducation en Ontario. Cela s'effectuera au moyen d'évaluations basées sur des données objectives, fidèles et pertinentes et de la publication en temps voulu de ces données et de la formulation de recommandations aux fins d'amélioration du système.

*La présente étude a été écrite sous contrat pour l'Office de la qualité et de la responsabilité en éducation (OQRE). Les opinions exprimées sont celles des auteurs et ne sont pas nécessairement celles de l'OQRE.*

© Imprimeur de la Reine pour l'Ontario 1999

Tous droits réservés. Aucune partie du présent document ne peut être reproduite, emmagasinée dans un système de recherche documentaire ou diffusée par moyen électronique, mécanique, sous forme de photocopie, d'enregistrement ou autre, sans l'accord préalable de l'Office de la qualité et de la responsabilité en éducation par l'intermédiaire de la chef des communications et du marketing.

Office de qualité et de la responsabilité en éducation
2, rue Carlton, bureau 1200
Toronto (Ontario) M5B 2M9

Téléphone : 1-888-327-7377

Site Web : www.eqao.com

ISBN 0-7778-9136-0

# Avant-propos

L'OQRE croit que les évaluations à grande échelle peuvent contribuer à apporter des changements constructifs au système d'éducation si elles amènent les éducatrices et les éducateurs, les parents et les élèves à réfléchir à ce qui se passe dans la classe et à en discuter.

Le mandat de l'OQRE est fondé sur deux engagements clés : la responsabilité et l'amélioration. Nous rendons compte aux parents, aux éducatrices et aux éducateurs et au public sur le rendement des élèves et la qualité de l'éducation dans le système scolaire de l'Ontario financé par les deniers publics. Nous veillons aussi à ce que ces renseignements servent de catalyseur pour améliorer l'apprentissage et l'enseignement.

L'une des façons de favoriser l'amélioration est de soutenir la recherche sur les facteurs qui ont une influence sur le rendement des élèves et la qualité de l'éducation. C'est ce que fait l'OQRE avec sa *Série d'études de recherche* qui fournit des analyses et des renseignements sur un vaste éventail de questions en rapport avec l'éducation.

Cette série a deux objectifs importants. Le premier est d'augmenter nos connaissances sur ce qui donne de bons résultats dans les classes, dans les écoles et dans les conseils scolaires. Le second est de combler les lacunes entre ce que nous savons et ce que nous faisons. En donnant aux éducatrices et aux éducateurs les connaissances, les occasions et la motivation nécessaires pour adapter et perfectionner leurs stratégies et leurs méthodes sur la base de données exactes et fidèles, la recherche peut avoir un effet positif et durable sur l'apprentissage des élèves. L'OQRE s'intéresse à toutes les formes de recherches systématiques qui favorisent le changement constructif et l'amélioration.

J'ai le plaisir de présenter la présente étude de la *Série d'études de recherche de l'OQRE.*

La directrice générale,

Joan M. Green
Août 1999

## TABLE DES MATIÈRES

Résumé ..... 1

Cadre méthodologique ..... 2
- Théorie de la généralisabilité ..... 2
- Classement et erreur de classement ..... 5

Application de l'analyse du classement et des erreurs de classement aux évaluations de l'OQRE ..... 7
- Collecte et sommaire des données ..... 7
- Le modèle statistique ..... 8
- Résultats ..... 10

Discussion et conclusion ..... 12

Références ..... 13

Les auteurs ..... 14

# Perspectives psychométriques pour l'OQRE : Théorie et applications de la généralisabilité

*Richard Wolfe*
*Institut d'études pédagogiques de l'Ontario*
*Université de Toronto*

*David Wiley*
*Université Northwestern*

*Ross Traub*
*Institut d'études pédagogiques de l'Ontario*
*Université de Toronto*

*La présente communication examine les perspectives psychométriques à partir desquelles définir et évaluer la généralisabilité des mesures et des rapports produits par l'Office de la qualité et de la responsabilité en éducation (OQRE). L'exactitude des mesures est exprimée en termes des composantes de la variabilité de la mesure : tâches, individus, correctrices et correcteurs et interactions connexes. Deux perspectives psychométriques sont utilisées pour analyser la généralisabilité des mesures des évaluations de l'OQRE. La première est l'analyse de la généralisabilité des sources de variation dans la notation des tests (c.-à-d. les composantes de la variance vraie et de la variance erronée); la seconde fait appel à des modèles de catégorisation vraie du rendement des élèves et à un cadre de probabilité pour définir le classement correct et le classement incorrect. Les données issues des évaluations de 3e année et de 9e année administrées en 1997–1998 par l'OQRE en français et en anglais sont examinées par une méthode d'analyse de structure latente qui estime les taux de classement correct et de classement incorrect des correctrices et des correcteurs. La probabilité de classement correct pour ces évaluations est estimée s'établir entre 70 pour 100 et 80 pour 100. Ce résultat est raisonnablement élevé et comparable à ce qui est observé dans d'autres évaluations à grande échelle du rendement.*

# Cadre méthodologique

## Théorie de la généralisabilité

La méthode la plus complète utilisée par les spécialistes des mesures éducatives pour analyser les erreurs dans la notation des exercices de rendement effectués par les élèves est basée sur la théorie de la généralisabilité. Cette méthode représente une extension et une expansion des méthodes utilisées en psychométrie dans le cadre de la théorie traditionnelle de fidélité, comme les cœfficients de fidélité fournis par la répétition des tests et les échanges de correctrices et de correcteurs. Elle s'appuie fortement sur la notion d'échantillonnage des tâches d'évaluation et des conditions de mesure à partir de populations plus importantes soumises à ces tests et dans ces conditions.

La théorie de la généralisabilité considère la fidélité de la mesure du point de vue de la généralisation de ces échantillons aux populations pertinentes. Les populations sont caractérisées par des paramètres qui représentent la variabilité dans la performance au test mesurée. Ces paramètres caractérisent deux types de variabilité :

1. la variabilité entre les individus dans les caractéristiques de la performance que l'on retrouve couramment dans les différentes tâches et dans les différentes conditions, c'est-à-dire les variations générales de la performance individuelle qui transcendent les circonstances particulières de la mesure;
2. la variation particulière de la performance individuelle qui ne reflète que des tâches et des conditions particulières.

Le premier type représente les stabilités de mesure désirées tandis que le deuxième représente les erreurs de mesure. Les études et les analyses de généralisabilité évaluent la magnitude de ces aspects de la variation de la performance en procédant à des analyses de la variance. Ces analyses décomposent la variation observée en composantes de variance représentant les personnes, les tâches et d'autres conditions de la mesure. Les magnitudes relatives et absolues des différentes composantes de la variance permettent de prédire l'exactitude qui serait obtenue avec des systèmes de mesure différents utilisant des conditions et des nombres différents des tâches étudiées. La théorie de la généralisabilité a été formulée pour la première fois par Cronbach, Gleser, Nanda et Rajaratnam (1972).

Dans les évaluations de performance, les erreurs de mesure sont le plus souvent associées aux tâches et à la notation. C'est-à-dire qu'elles apparaissent dans les variations de la notation d'un individu pour des tâches similaires et dans la variation des notations pour une même tâche. Plus précisément, ces erreurs sont évaluées relativement à la variation :

- de la notation de différents exercices considérés comme interchangeables aux fins de l'évaluation;
- des notations assignées à une seule tâche également considérées comme interchangeables aux fins de l'évaluation.

Plus ces variations sont importantes, plus grande est l'erreur de mesure et plus faible l'exactitude. Dans la plupart des évaluations de performance, les exercices notés dans un contexte de notation particulier pour une matière particulière et une classe particulière sont habituellement considérés comme interchangeables, particulièrement lorsque les correctrices et correcteurs sont formés de la même façon de manière à atteindre le même niveau de compétence.

Les données pour une analyse de généralisabilité de ces types d'erreurs sont souvent obtenues à partir d'une situation où chaque membre d'un groupe de personnes soumises à l'évaluation exécute chaque tâche en une suite constante et où chaque réponse est notée par le même ensemble de correctrices et de correcteurs. Une analyse des données produites dans ce type de structure, lorsqu'il y a plus d'une tâche dans la série, produit les estimations des composantes de variabilité illustrées au Tableau 1.

## Cadre méthodologique

### Théorie de la généralisabilité

Tous les types d'erreurs énumérés dans le Tableau 1 contribuent à l'erreur de mesure. Ceci est vrai quel que soit le nombre de tâches effectuées par la répondante ou le répondant dans une série de tâches interchangeables ou le nombre de notations interchangeables assignées à une tâche. Cependant, suivant la conception de l'évaluation, p. ex., le nombre exact de tâches effectuées par une répondante ou un répondant ou le nombre de correctrices et de correcteurs, les effets des différentes composantes varieront. Si seulement une tâche d'un type donné est effectuée par les répondantes et répondants ou si seulement une notation est obtenue pour une tâche, certaines de ces composantes ne seront pas estimables.

On peut faire remonter la plupart de ces types d'erreurs à deux sources générales : les différences dans la mesurabilité des tâches et les différences entre les correctrices et correcteurs dans l'évaluation de performance. Comme les tâches sont sélectionnées, et donc fixées, avant que la notation ait lieu, les différences entre les tâches peuvent être conçues comme antérieures aux effets de la notation et indépendantes de ces effets. Par ailleurs, les effets de la notation ne peuvent se produire que dans le contexte des tâches assignées. De ce point de vue, il est logique de définir les composantes d'erreurs associées aux différences entre les tâches (effets des tâches) comme reliées aux tâches mais non à la notation et de définir les effets de la notation comme reliées à la notation, qu'elles soient ou non reliées aux tâches. C'est-à-dire que les résultats d'une tâche particulière différeront habituellement de façon aléatoire de ceux qui seraient obtenus après une évaluation de performance à un large éventail de tâches interchangeables. La notation, en revanche, dépend de chaque tâche : les correctrices et les correcteurs ne notent que des tâches précises. De ce point de vue, les composantes de la variance d'erreur du Tableau 1 seraient classées comme indiqué dans le Tableau 2.

L'erreur due aux différences entre les tâches à accomplir, à l'exclusion de la notation, s'exprime en fidélité comme :

$r_t = \sigma^2_p / (\sigma^2_p + \sigma^2_t + \sigma^2_{pt})$.

L'erreur due à la notation, une fois les tâches fixées (de sorte que la notation vraie d'une tâche particulière est la notation

**TABLEAU 1 Types de variation**

| Variance de la composante | Type de variation |
|---|---|
| Variation vraie | |
| $\sigma^2_p$ | Variation générale « vraie » entre les personnes dont on évalue le rendement |
| Variation aléatoire | |
| $\sigma^2_s$ | Variation entre les correctrices et correcteurs en indulgence moyenne |
| $\sigma^2_t$ | Variation entre les tâches en difficulté moyenne |
| $\sigma^2_{ps}$ | Variation entre les correctrices et les correcteurs et les personnes en indulgence envers les individus |
| $\sigma^2_{pt}$ | Variation entre les tâches et les personnes en difficulté de la tâche pour les individus |
| $\sigma^2_{st}$ | Variation entre les correctrices et les correcteurs et les tâches en difficulté de la tâche notée |
| $\sigma^2_{pst}$ | Variation entre les correctrices et les correcteurs, les tâches et les personnes en difficulté de la tâche et en indulgence envers les individus |

observée une fois l'erreur ajoutée) s'exprime en fidélité comme :

$$r_s = (\sigma^2_p + \sigma^2_t + \sigma^2_{pt}) / (\sigma^2_p + \sigma^2_t + \sigma^2_{pt} + \sigma^2_s + \sigma^2_{ps} + \sigma^2_{st} + \sigma^2_{pst}).$$

Comme la fidélité totale, reflétant les deux sources d'erreurs, est :

$$r_{tot} = \sigma_p^2 / (\sigma_p^2 + \sigma_t^2 + \sigma_{pt}^2 + \sigma_s^2 + \sigma_{ps}^2 + \sigma_{st}^2 + \sigma_{pst}^2),$$

on a donc :

$$r_{tot} = r_t r_s.$$

C'est-à-dire que la fidélité totale est le produit de la fidélité liée à l'erreur due aux tâches et de la fidélité liée à l'erreur due à la notation.

Bien que toutes les composantes ci-dessus soient présentes dans le processus d'évaluation de l'OQRE, la conception actuelle de l'évaluation impose des limites quant à celles qui peuvent être évaluées. Des études de recherche pourraient être faites pour dégager les composantes qui peuvent effectivement être intégrées à la principale étape de recueil des données et incluses dans l'analyse subséquente.

**TABLEAU 2 Composantes d'erreurs pour les tâches et la notation**

| Source | Composantes d'erreurs |
|---|---|
| Tâche | $\sigma^2_t$, $\sigma^2_{pt}$ |
| Notation | $\sigma^2_s$ , $\sigma^2_{ps}$ , $\sigma^2_{st}$ , $\sigma^2_{pst}$ |

## Cadre méthodologique

### Classement et erreur de classement

L'analyse des erreurs de mesure suivant la méthode de généralisabilité traite les réponses analysées comme s'il était possible de leur assigner des valeurs sur une échelle continue. Une autre méthode assigne les réponses à une ou plusieurs catégories discrètes et ordonnées. De ce point de vue, l'erreur de mesure vient de l'assignation de notations à des catégories incorrectes. La façon naturelle de quantifier ces erreurs de classement est en termes de probabilité de classement correct ou d'erreur de classement, c'est-à-dire de probabilités comme : P(i|j) = la probabilité d'une réponse étant assignée à la catégorie i alors que j est en fait la catégorie correcte.

L'erreur de classement peut être due à chacun des types d'erreurs énumérés dans les Tableaux 1 et 2. Plus particulièrement, si un individu était vraiment dans la catégorie j, une tâche t pourrait avoir la probabilité P(i|j;t) de classer cette personne dans la catégorie i. Cette probabilité d'erreur de classement liée à une tâche émanerait des mêmes sources que celles qui sont résumées par $\sigma^2_t$ et $\sigma^2_{pt}$. Une moyenne de ces probabilités d'erreurs de classification de toutes les tâches, à savoir la moyenne$_t${P(i|j;t)}, résume la relation générale parmi les tâches entre les catégories vraies et observées.

Les relations d'erreurs de classement peuvent être présentées sous forme de matrice d'erreur de classement, comme illustré dans le Tableau 3.

Ces matrices convertissent les valeurs vraies en valeurs observées. C'est-à-dire que si $y_p$ est un vecteur représentant le classement d'un individu p dans une catégorie observée (c.-à-d. que le vecteur = 1 pour la catégorie observée et 0 pour chaque autre catégorie), et si $x_p$ est le vecteur vrai correspondant, alors $E_m(y_p) = Mx_p$, où $E_m$ est l'attente pour tout le processus d'erreur de classement (performance de la tâche ou notation). De même, $E_p[E_m(y_p)] = E_p[Mx_p] = ME_p[x_p]$, où $E_p$ est l'attente pour l'ensemble de la population. On remarque que $E_p[x_p] = \pi$, les proportions vraies des catégories dans la population, et $E_p[E_m(y_p)] = p$, les proportions observées des catégories dans la population. C'est ainsi que $p = M\pi$. C'est-à-dire que les matrices d'erreurs de classement transforment les distributions vraies pour l'ensemble des catégories de groupes de répondantes et de répondants en distributions observées pour l'ensemble des catégories de tâches ou de notations. Si nous indexons ces valeurs pour une tâche particulière, soit t, alors $p_t = M_t\pi$.

Si l'erreur de classement due à la notation de la tâche t est $M_{ts}$, alors $p_{ts} = M_{ts}p_t = M_{ts}M_t\pi$, si bien que la matrice d'erreur de classement de la performance vraie après sélection et la notation des tâches est le produit de l'erreur de classement liée à la matrice de sélection de la tâche et de la matrice d'erreur liée au classement de la notation. Il faut également noter que $\Sigma_t(p_{ts}) = \Sigma_t(M_{ts}M_t\pi) = [\Sigma_t(M_{ts}M_t)]\pi$, si bien que la

**TABLEAU 3 Formulation de la matrice d'erreurs de classement M**

| | Catégorie vraie | | | |
|---|---|---|---|---|
| **Catégorie observée** | 1 | 2 | 3 | 4 |
| 1 | P(1\|1) | P(1\|2) | P(1\|3) | P(1\|4) |
| 2 | P(2\|1) | P(2\|2) | P(2\|3) | P(2\|4) |
| 3 | P(3\|1) | P(3\|2) | P(3\|3) | P(3\|4) |
| 4 | P(4\|1) | P(4\|2) | P(4\|3) | P(4\|4) |
| **Total** | 1,00 | 1,00 | 1,00 | 1,00 |

somme et la moyenne des erreurs de classement pour l'ensemble des tâches représentent les erreurs de classement pour la somme et la moyenne des proportions observées.

Chaque entrée diagonale dans la matrice d'erreur de classement représente la probabilité de classement correct, c'est-à-dire que la tâche d'un individu dans une catégorie vraie particulière sera notée dans cette catégorie. Ces entrées diagonales représentent l'exactitude ou la fidélité de la mesure. (Il faut noter que les indices de fidélité classiques, tels que décrits ci-dessus pour la théorie de la généralisabilité, sont des rapports de corrélation et non des probabilités. Nous trouvons les probabilités d'analyse de structure latente plus interprétables, surtout quand des échelles catégoriques sont utilisées comme dans les évaluations de l'OQRE. Bien que l'étendue des probabilités et des rapports de corrélation soit la même, zéro à un, l'échelle ne l'est pas, si bien que les deux ne sont pas directement comparables. Nous présentons une façon de transformer les indices classiques de fidélité en probabilités plus loin dans cet article.)

# Application de l'analyse des classements et des erreurs de classement aux évaluations de l'OQRE

Dans les évaluations de 3e année et de 9e année de l'OQRE, nous devons considérer chaque élève comme ayant effectué une seule tâche, à savoir la procédure d'évaluation tout entière. Compte tenu des priorités et des objectifs généraux de l'OQRE, il n'était pas faisable de répéter les tâches dans aucun des tests, c'est-à-dire de séparer des réponses d'élèves pouvant être notées séparément et être considérées comme interchangeables. Cela signifie que les composantes de variance qui dépendent de la tâche (variance de la tâche, tâche par interaction des élèves, tâche par interaction des correctrices et correcteurs et tâche par élève par interaction des correctrices et correcteurs) ne pouvaient pas être évaluées. Nous devons donc nous concentrer sur les variances et les cœfficients de fidélité dus aux correctrices et correcteurs.[1]

## COLLECTE ET SOMMAIRE DES DONNÉES

Les données de la présente analyse dérivent d'un système de répétition de la correction utilisé par l'OQRE. Une fois qu'entre 500 à 1 000 cahiers ont été corrigés, on procède à une analyse rapide de la distribution provisoire des notes. Un échantillon stratifié de cahiers est choisi avec une distribution relativement uniforme sur l'étendue des notes. La note générale ou moyenne est utilisée dans cette sélection. Pour l'évaluation provinciale en anglais, qui est administrée à plus de 100 000 élèves, environ 200 cahiers sont choisis aux fins de calibrage. Pour l'évaluation en français et lors des mises à l'essai (c.-à.-d. l'évaluation de 9e année) quelque 100 cahiers sont choisis aux fins de calibrage. Une fois sélectionnés, les cahiers sont réinsérés aussi souvent que possible dans le flot des cahiers à corriger sans indication visible qu'il s'agit d'une répétition. Comme il y a un grand nombre de cahiers à corriger et un grand nombre de correctrices et de correcteurs, il est très rare qu'un cahier de calibrage soit noté deux fois par la même personne. Suivant la rapidité de la correction et le nombre de cahiers et de correctrices et correcteurs, chaque cahier de calibrage est corrigé à nouveau de trois à six fois.

Les données des corrections répétées sont résumées, séparément pour chaque évaluation et pour chaque note qui doit être analysée, en tabulation en croix. On trouvera un exemple de tabulation en croix au Tableau 4. Dans la forme A de l'exemple, un cahier choisi au hasard corrigé par deux correctrices ou correcteurs sélectionnés au hasard a 21,15 pour 100 de chances d'être noté deux fois au niveau 2 et 12,61 pour 100 de chances d'être noté une fois au niveau 2 et une fois au niveau 3. Les figures correspondantes pour la forme B sont 18,33 pour 100 et 13,61 pour 100.

La distribution générale des notes peut être déterminée à partir de la tabulation en croix en ajoutant les chiffres en diagonale à la moitié de la somme des chiffres hors diagonale (puisque la tabulation en croix enregistre les notes en groupes de deux). Les distributions sont 14,98 pour 100, 34,95 pour 100 et 16,43 pour 100 pour les niveaux 1, 2. 3 et 4 pour la forme A; et 19,45 pour 100, 34,58 pour 100, 32,08 pour 100 et 13,90 pour 100 pour la forme B. Bien que le même nombre de cahiers aient été choisis aux fins de réinsertion dans chacun des quatre niveaux, après la première correction des cahiers, on assiste à une certaine régression

[1] Dans toutes les évaluations de l'OQRE, il y a des notes séparées pour les principales compétences et domaines, et pour les sujets ou matières. Elles fournissent des évaluations différenciées des élèves mais ne sont pas remplaçables ou interchangeables aux fins de détermination de la fidélité. Dans plusieurs évaluations et dans les mises à l'essai, il y a eu des notations « analytiques » où des notes partielles ont été appliquées aux différentes questions et aux différents segments du travail des élèves. Celles-ci corrélaient bien avec les notes globales. Les questions à réponse choisie et à réponse courte soumises à des échantillons d'élèves corrélaient également bien avec les notes globales. Quoique ceci augure bien de la fidélité des notes globales (les corrélations ne pourraient pas exister sans fidélité au moins à ce niveau), il n'est pas possible de quantifier la fidélité à partir de ces connexions parce que les méthodes pour ce type d'analyse n'ont pas encore été développées.

vers le milieu dans les notations moyennes, c'est-à-dire que les pourcentages de niveaux 1 et 4 sont diminués alors que ceux des niveaux 2 et 3 sont augmentés.

## LE MODÈLE STATISTIQUE

Les sommaires des données dépendent de la conjonction entre les données de distribution des notes et les taux d'accord entre les correctrices et correcteurs. Pour désagréger ces effets, nous appliquons une analyse de structure latente comme illustré à la figure 1. (Pour l'explication initiale de l'analyse de structure latente, voir Lazarsfeld et Henry, 1968.)

Pour comprendre l'analyse de structure latente, nous devons considérer qu'aux fins d'application réelle, nous ne savons pas quelle est la notation « vraie » d'un cahier. Nous pourrions essayer de la dériver d'un consensus d'experts mais, même dans ce cas, nous avons peu de chances d'obtenir un accord à 100 pour 100. En l'absence d'un classement « vrai », nous ne pouvons pas calculer directement la proportion de classement correct et incorrect. En revanche, les proportions de l'accord, comme indiqué au Tableau 4, dépend à la fois de la distribution des notes vraies hypothétiques (les catégories latentes de cahiers) et des probabilités de classement des correctrices et correcteurs (voir le Tableau 3).

L'analyse de la catégorie latente propose un modèle simple pour le processus de notation. D'abord, les cahiers ont un classement vrai, bien qu'inconnu, latent (ou vrai) et commun à toutes les correctrices et les correcteurs. Ensuite, quand une correctrice ou un correcteur assigne une note à un cahier, il s'agit d'un classement correct avec une certaine probabilité. Avec la probabilité complémentaire, la correctrice ou le correcteur relève ou rabaisse la note d'un niveau. Au niveau 1, (le niveau le plus bas), on considère qu'il ne peut y avoir qu'une note correcte ou relevée, notes rabaissée et correcte étant semblables et au niveau 4 (le niveau le plus élevé), il ne peut y avoir qu'une note correcte ou rabaissée, notes relevée et correcte étant semblables. Aux autres niveaux, il y a des probabilités constantes mais différentiées de relèvement et de rabaissement.

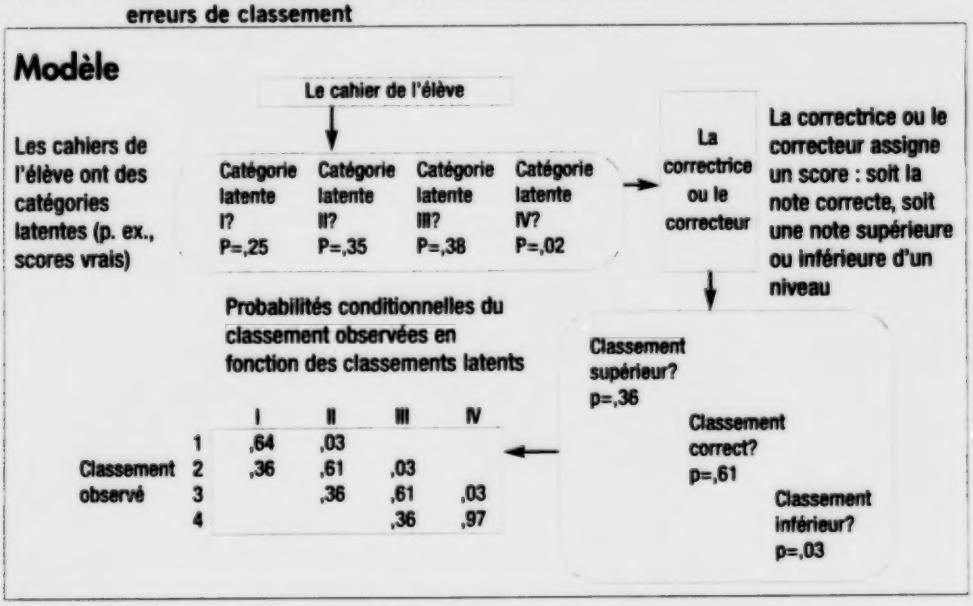
**TABLEAU 4 Exemple de résumés de données de classement et d'erreurs de classement**

| Forme A Niveau | Niveau 1 | Niveau 2 | Niveau 3 | Nvieau 4 |
|---|---|---|---|---|
| 1 | 7,70 | | | |
| 2 | 14,01 | 21,15 | | |
| 3 | 0,56 | 12,61 | 18,21 | |
| 4 | 0 | 0,98 | 17,60 | 7,14 |
| **Forme B Niveau** | | | | |
| 1 | 10,28 | | | |
| 2 | 15,83 | 18,33 | | |
| 3 | 2,50 | 13,61 | 17,08 | |
| 4 | 0 | 3,06 | 13,89 | 5,42 |

*Remarque* : Dans chaque tabulation en croix, la somme des entrées est 100 et elles donnent une estimation de la distribution des notes comme si un cahier choisi au hasard était noté par deux correctrices ou correcteurs choisis au hasard.

## Application de l'analyse des classements et des erreurs de classement aux évaluations de l'OQRE

**FIGURE 1 Modèle de catégorie latente pour estimer la fidélité de la notation comme erreurs de classement**


Modèle
Le cahier de l'élève
Les cahiers de l'élève ont des catégories latentes (p. ex., scores vrais)
Catégorie latente I? P=,25
Catégorie latente II? P=,35
Catégorie latente III? P=,38
Catégorie latente IV? P=,02
La correctrice ou le correcteur
La correctrice ou le correcteur assigne un score : soit la note correcte, soit une note supérieure ou inférieure d'un niveau
Probabilités conditionnelles du classement observées en fonction des classements latents
Classement observé
I II III IV
1 ,64 ,03
2 ,36 ,61 ,03
3 ,36 ,61 ,03
4 ,36 ,97
Classement supérieur? p=,36
Classement correct? p=,61
Classement inférieur? p=,03

## Données attendues

| | I | II | III | IV |
|---|---|---|---|---|
| 1 | 7,7 | | | |
| 2 | 14,0 | 21,2 | | |
| 3 | ,6 | 12,6 | 18,2 | |
| 4 | | 1,0 | 17,6 | 7,1 |

Tabulation en croix des moyennes théoriques définies par les paramètres (la somme est 100 pour 100)

## Données observées

| | I | II | III | IV |
|---|---|---|---|---|
| 1 | 10,1 | | | |
| 2 | 12,4 | 16,4 | | |
| 3 | ,6 | 16,7 | 19,0 | |
| 4 | | ,7 | 17,0 | 7,1 |

Tabulation en croix des moyennes observées, calculées à partir des données calibrées (la somme est 100 pour 100)

*Remarque :* Dans l'analyse statistique, nous essayons de trouver les paramètres d'un modèle avec lequel les données observées ont le plus de chance de coïncider, c'est-à-dire qui soit le plus proche des données attendues, compte tenu des paramètres.

## Application de l'analyse des classements et des erreurs de classement aux évaluations de l'OQRE

Les proportions des cahiers dans les catégories vraies sont des paramètres du modèle à estimer. Les probabilités de notation correcte, de relèvement et de rabaissement de la note représentent trois autres paramètres. Cet ensemble de paramètres est estimé à partir du sommaire des données par la méthode de la probabilité maximale. Cela signifie qu'à l'issue d'une itération numérique, des valeurs de paramètres sont trouvées qui rendent les proportions sommaires observées aussi probables que possible.

De retour à la Figure 1, nous voyons qu'un diagramme du modèle est présenté dans la rangée du haut avec les paramètres pour les catégories latentes et le classement ou l'erreur de classement tel qu'estimé à partir des données dans la rangée du bas. La rangée du milieu montre la valeur attendue des données compte tenu des paramètres et l'on peut voir qu'il s'agit d'une proche approximation des données.

Ce modèle particulier de catégorie latente est un cas spécial du modèle présenté dans la section précédente. La simplification est nécessaire à cause des tailles relativement faibles des échantillons. Le modèle semble être adéquat pour correspondre aux données, en partie parce qu'il y a très peu de cas où les correctrices et correcteurs diffèrent de plus d'un niveau dans leur jugement.

### RÉSULTATS

Les résultats de l'application du modèle de catégorie latente aux évaluations de 1997–1998 sont présentés dans les Tableaux 5 et 6. Comme on l'a fait remarquer plus tôt, on peut considérer les probabilités de classement vrai comme une bonne mesure sommaire de la fidélité, c'est-à-dire de la probabilité qu'une ou un élève sera noté « correctement » par une correctrice ou un correcteur. Ces probabilités s'établissent entre 70 et 80 pour 100. Ceci est raisonnablement élevé et comparable aux niveaux que l'on trouve habituellement dans d'autres évaluations à grande échelle (voir, par exemple, Dunbar, Koretz et Hoover, 1991). De plus, ces taux de classement sont comparables à ce qui serait obtenu avec les tests traditionnels à question à réponse choisie (voir la section suivante).

**TABLEAU 5 Analyse de catégorie latente pour les niveaux assignés en 1997–1998 en français**

| Niveau | LC0 | LC1 | LC2 | LC3 | LC4 | bas | haut | vrai |
|---|---|---|---|---|---|---|---|---|
| Français 3e année Lecture | 0,10 | 0,05 | 0,30 | 0,55 | 0,00 | 0,13 | 0,12 | 0,74 |
| Français 3e année Écriture | 0,10 | 0,06 | 0,33 | 0,53 | 0,00 | 0,09 | 0,16 | 0,74 |
| Français 3e année Math | 0,12 | 0,05 | 0,29 | 0,46 | 0,08 | 0,12 | 0,08 | 0,79 |
| Français 9e année Math | 0,07 | 0,13 | 0,40 | 0,38 | 0,02 | 0,25 | 0,05 | 0,69 |

**TABLEAU 6 Analyse de catégorie latente pour les niveaux assignés en 1997–1998 en anglais**

| Niveau | LC0 | LC1 | LC2 | LC3 | LC4 | bas | haut | vrai |
|---|---|---|---|---|---|---|---|---|
| Anglais 3e année Lecture | 0,10 | 0,08 | 0,35 | 0,42 | 0,05 | 0,11 | 0,07 | 0,81 |
| Anglais 3e année Écriture | 0,09 | 0,09 | 0,46 | 0,36 | 0,00 | 0,06 | 0,20 | 0,74 |
| Anglais 3e année Math | 0,09 | 0,11 | 0,41 | 0,38 | 0,00 | 0,08 | 0,18 | 0,73 |
| Anglais 9e année Math | 0,12 | 0,15 | 0,44 | 0,28 | 0,01 | 0,02 | 0,26 | 0,72 |

## Application de l'analyse des classements et des erreurs de classement aux évaluations de l'OQRE

### Taux de classement dans les tests conventionnels à réponse choisie

Pour des raisons de contenu et de validité corrélative, les évaluations de l'OQRE utilisent essentiellement des tâches à exécuter qui peuvent être mesurées de façon globale. Il est intéressant de comparer les cœfficients de fidélité qui pourraient être obtenus si l'OQRE utilisait à la place des évaluations de performance des tests traditionnels. Les élèves pourraient être classés en quatre niveaux (pour correspondre aux niveaux de rendement de l'OQRE) en divisant la distribution des notes à un test gradué conventionnel comme un test à réponse choisie, suivant des seuils de note appropriés. En utilisant la théorie classique de la fidélité des tests, on peut estimer les probabilités de classement correct qui seraient obtenues par cette méthode.

On propose à titre d'exemple qu'un test gradué ait donné lieu à une distribution de niveaux finals similaires aux niveaux de l'évaluation de 3e année administrée par l'OQRE en 1997–1998, avec 10 pour 100 pour le niveau 1, 40 pour 100 pour le niveau 2, 45 pour 100 pour le niveau 3 et 5 pour 100 pour le niveau 4. (Ces chiffres représentent une distribution approximative moyenne à travers les matières et les langues de l'OQRE).

Considérons des tests gradués avec des cœfficients de fidélité classiques (ρ) de 0,70, 0,80 et 0,90, correspondant à une exactitude moyenne, élevée et très élevée pour les tests éducationnels commerciaux en mathématiques et en français. Pour chaque test, nous pourrions calculer les probabilités d'assignation de note au niveau correct pour les élèves qui sont au centre des niveaux 1 à 4, comme indiqué au Tableau 7.

Par exemple, au Tableau 7, le chiffre 73 pour le niveau 2, avec une fidélité de 0,70 indique qu'il y a 73 pour 100 de chance qu'une ou un élève au vrai centre du niveau 2 de ce test (ce qui signifie, dans ce cas, au 30e percentile de la distribution vraie des notes) sera classé au niveau 2. (Une analyse plus détaillée indique que les élèves ont 6 pour 100 de chance d'être classés au niveau 1, 21 pour 100 de chance d'être classés au niveau 3 et pratiquement aucune chance d'être classés au niveau 4.)

Nous voyons que les probabilités de classement correct sont plus élevées pour les niveaux 2 et 3 parce qu'il s'agit de larges catégories et que nous considérons seulement les élèves qui se trouvent au centre. Pour les élèves plus proches du seuil des niveaux, les probabilités de classement correct seraient plus faibles. Les chances de classement correct sont aussi substantiellement plus faibles dans les niveaux extérieurs, 1 et 4, parce que ces niveaux sont plus étroits et parce qu'il y a une régression générale vers le centre.

La leçon à tirer est que les tests conventionnels à réponse choisie, même s'ils ont une haute fidélité, produiront des erreurs de classement substantielles lorsqu'ils seront utilisés pour assigner des niveaux aux élèves. En outre, les tests à réponse choisie sont jugés présenter des défauts considérables en matière de validité et de partialité. Par ailleurs, les taux de classement correct par une seule correctrice ou un seul correcteur tels que produits dans les évaluations de performance de l'OQRE sont d'une magnitude comparable.

**TABLEAU 7 Probabilité de classement correct des élèves au centre du niveau pour les tests avec des cœfficients de fidélité différents**

| Fidélité | Niveau 1 | Niveau 2 | Niveau 3 | Niveau 4 |
|---|---|---|---|---|
| 0,70 | 57 | 73 | 80 | 50 |
| 0,80 | 66 | 82 | 88 | 60 |
| 0,90 | 81 | 94 | 96 | 75 |

# Discussion et conclusion

On fait parfois appel aux cœfficients de fidélités inter-tâches et inter-correcteurs pour ce type de test et de correction. Cependant, une analyse de la fidélité inter-tâches n'est pas faisable ici parce qu'il n'y a qu'une tâche par élève. Aux fins de la présente analyse, nous nous sommes donc attachés à la fidélité inter-correcteurs. Un échantillon de cahiers avec une distribution uniforme sur l'échelle de notation a été réinséré entre trois et six fois dans l'ensemble des cahiers à corriger. Un cahier n'a pratiquement jamais été corrigé par la même correctrice ou le même correcteur. Ce qui était calculé dans le modèle présenté ici est l'exactitude de la note qui était assignée à chaque cahier comme indiqué par la probabilité de classement correct. (Ce calcul est clairement différent de la probabilité que deux correctrices ou correcteurs coïncident dans leur notation. Bien que l'on puisse calculer la probabilité que deux correctrices ou correcteurs coïncident, dans ce modèle cette probabilité inclurait des cas où les correctrices et les correcteurs coïncident dans leurs erreurs.)

Il est important de souligner que les distributions de structure latente ne sont pas les mêmes que la distribution originale des notes. On peut présumer qu'elles donnent une meilleure image de la distribution des classements « vrais » de ces élèves parce que les distributions observées sont déformées par les erreurs de classement. Il faut remarquer que l'échantillon de cahiers utilisé dans l'exercice de calibration a été choisi pour représenter une distribution plus ou moins uniforme sur l'échelle de notation, si bien que les proportions de structure latente peuvent ne pas refléter la population tout entière. En tous cas, on peut voir que le nombre de cas dans la catégorie latente 4 est très faible.

Ces analyses nous permettent de commencer une évaluation de la suffisance de la fidélité des mesures et des rapports de l'OQRE. Généralement parlant, les probabilités de classement correct sont raisonnablement élevées. Ces taux sont comparables à ce que l'on trouve dans d'autres évaluations à grande échelle et ils sont équivalents aux taux de classement qui seraient obtenus dans des tests plus traditionnels à réponse choisie, qui sont généralement considérés comme présentant un degré très élevé de fidélité.

# Références

Cronbach, L. J., Gleser, G. C., Nanda, H. et Rajaratnam, N. *The dependability of behavioral measurements: Theory of generalizability of scores and profiles.* New York : Wiley, 1972.

Dunbar, S.B., Koretz, D.M. et Hoover, H.D. Quality control in the development and use of performance assessments. *Applied Measurement in Education*, 4(4), 1991, p. 289-303

Lazarsfeld, P. F. et Henry, N. W. *Latent structure analysis.* Boston : Houghton Mifflin, 1968.

# Les auteurs

Richard Wolfe est professeur agrégé dans le programme de mesure et d'évaluation du Département du curriculum, de l'enseignement et de l'apprentissage de l'Institut d'études pédagogiques de l'Ontario de l'Université de Toronto. Son champ de recherches comprend la conception et l'analyse des évaluations à grande échelle.

David Wiley est professeur, en congé, dans le programme des sciences de l'apprentissage, et ancien doyen de l'École d'éducation et de politique sociale de l'Université Northwestern. Ses recherches portent sur les statistiques et sur la théorie du curriculum, de l'apprentissage et de l'évaluation.

Ross Traub est professeur dans le programme de mesure et d'évaluation du Département du curriculum, de l'enseignement et de l'apprentissage de l'Institut d'études pédagogiques de l'Ontario de l'Université de Toronto. Ses recherches portent sur la psychométrie théorique et appliquée.